Nucleic Acids Research, 2007, Vol. 35, Web Server issue ***W429–W432***
doi:10.1093/nar/gkm256

Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server

Lukas Käll[1,*], Anders Krogh[2] and Erik L.L. Sonnhammer[1]

[1]Center for Genomics and Bioinformatics, Karolinska Institutet, S-17177 Stockholm, Sweden and
[2]Department of Molecular Biology, University of Copenhagen, Ole Maaloes Vej 5, 2200 Copenhagen, Denmark

Received January 26, 2007; Revised March 22, 2007; Accepted April 8, 2007

ABSTRACT

When using conventional transmembrane topology and signal peptide predictors, such as TMHMM and SignalP, there is a substantial overlap between these two types of predictions. Applying these methods to five complete proteomes, we found that 30–65% of all predicted signal peptides and 25–35% of all predicted transmembrane topologies overlap. This impairs predictions of 5–10% of the proteome, hence this is an important issue in protein annotation.

To address this problem, we previously designed a hidden Markov model, Phobius, that combines transmembrane topology and signal peptide predictions. The method makes an optimal choice between transmembrane segments and signal peptides, and also allows constrained and homology-enriched predictions.

We here present a web interface (http://phobius.cgb.ki.se and http://phobius.binf.ku.dk) to access Phobius.

INTRODUCTION

Traditional transmembrane topology predictors often predict signal peptides as transmembrane segments, and vice versa signal peptide predictors often predict N-terminal transmembrane segments as signal peptides. This fact is often overlooked when testing prediction methods, and is the main cause for very different test results. A frequent advice how to circumvent the problem of these cross-predictions is to remove predicted signal peptides before predicting transmembrane proteins (1), or to remove proteins with transmembrane segments when predicting signal peptides (2). However, as the number of errors due to cross predictions is roughly the same for the two kinds of predictors (3), the gain will be as high as the loss by such approaches.

To resolve the ambiguities we have, in a previous study, designed a hidden Markov model, Phobius, containing submodels for both signal peptides and transmembrane segments (see Figure 1). We obtain better discrimination by forcing the predictor to chose between the two types of features. A benchmark (3) showed that false classifications of signal peptides were reduced from TMHMM's (4) 26 to 4% and false classifications of transmembrane helices were reduced from SignalP 2.0's (5) 19 to 8%. An advantage is that the method even increased the high accuracy of TMHMM in predicting pure transmembrane topologies from 44.5 to 53.9% correctly predicted topologies. Since this benchmark, a new version SignalP 3.0 (6) has been published. Its false positive rate on transmembrane proteins is however as high as before. On the same set of transmembrane proteins without signal peptides used in the previous benchmark, SignalP 3.0 produces false predictions on 21% (52 of 247) of the test sequences.

Here, we present an overlap analysis between signal peptides predictions and transmembrane segment predictions done by conventional predictors on five proteomes. We also give a description of the Phobius web interface.

WHOLE PROTEOME OVERLAP ANALYSIS

To investigate how large a problem the overlap between predictions between conventional signal peptide predictors and transmembrane topology predictors are at whole proteome level, we tried to annotate five different proteomes using a combination of SignalP 3.0 (6) and TMHMM 2.0 (4). The results are given in Table 1.

*To whom correspondence should be addressed. Tel: +1 206 616 5021; Fax: +1 206 685 7301; Email: lukall@u.washington.edu
Present addresses:
Lukas Käll, Department of Genome Sciences, University of Washington, Seattle WA, USA
Erik L.L. Sonnhammer, Stockholm Bioinformatics Center, Stockholm University, Stockholm, Sweden

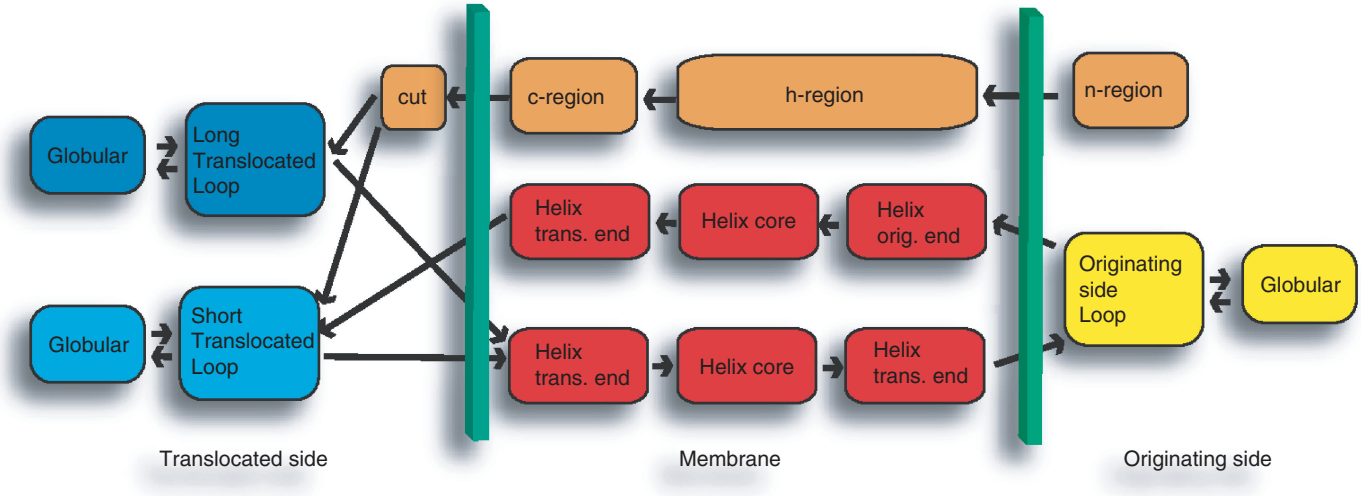

Figure 1. The Phobius model. The model comprise submodels for signal peptides, transmembrane helices, cytoplasmic loops and two different submodels for non-cytoplasmic loops.

Table 1. Overlapping signal peptide and transmembrane segments in whole proteome predictions by conventional predictors

	Overlap/TM predictions	Overlap/SP predictions	Overlap/All sequences
Human (CCDS)	30% (1113/3720)	32% (1113/3485)	7.6% (1113/14663)
C. elegans	34% (2587/7694)	41% (2587/6328)	9.6% (2587/26032)
S. cerevisiae	26% (377/1468)	48% (377/787)	5.6% (377/6680)
E. coli K12	26% (271/1039)	39% (271/698)	6.4% (271/4243)
B. subtilis	32% (358/1133)	63% (358/565)	8.7% (358/4105)

The proteomes of five different species were annotated with SignalP-NN 3.0 and TMHMM 2.0. Predictions were counted as overlapping if a part of a potential signal peptide as predicted by SignalP also was predicted as a transmembrane helix by TMHMM. In such cases, at least one of the prediction methods is wrong. The overlapping predictions are expressed as fractions of all predicted transmembrane proteins, all signal peptide predictions and the number of sequences in the proteome. The SignalP-NN predictions were carried out using the optional 70 residue truncation and the correct organism group, and were counted as predicted signal peptides if the D-score was over threshold. The proteomes of *Caenorhabditis elegans* and *cerevisiae* and the consensus coding sequences of the human proteome were downloaded from Ensembl and the proteomes of *Escherichia coli* K12 and *Bacillus subtilis* were downloaded from NCBI's web site.

We found that 5–10% of all the proteins have predicted transmembrane segments that overlap predicted signal peptides. Since only one of the methods can be correct, this casts doubt on 30–65% of all predicted signal peptides and 25–35% of all predicted transmembrane topologies. Both predictions are roughly equally frequent in a proteome, and their false positive rates are more or less the same, hence we cannot tell which method is correct based on SignalP and TMHMM predictions in these cases. Phobius thus solves a problem that other signal peptide predictors and transmembrane topology predictors cannot handle.

DESCRIPTION OF WEB INTERFACE

The Phobius web server provides an easy and accurate mean to predict signal peptides and transmembrane topology from an amino acid sequence. The sequences should be submitted in fasta format, preferably uploaded as a file. The predictions are given either in 'short'—single line text output or 'long'—UniProt feature table styled output (see Figure 2).

All predictions made by the Phobius server can optionally be accompanied by a posterior label (location) probability plot. The posterior label probability is the probability for a location (cytoplasm, non-cytoplasm, membrane or signal peptide) of a residue given the whole sequence (see Figure 2). Note that the posterior probability plot is not a prediction in itself. The pattern of the plot might even deviate from the prediction, which would be a sign of uncertainty in the prediction.

In 'normal prediction' mode as well as in the 'constrained prediction' mode described below, sequences are decoded with the 1-best algorithm (7).

CONSTRAINED PREDICTION

The accuracy of the predictions can be greatly improved if we can include information about the location of a part of the sequence in a constrained prediction (8). Typically we could have experimental data at hand from reporter fusions (9), antibody experiments, or have knowledge of the location due to functional requirements of a site (10). The Phobius web server provides a service to let the user

Phobius prediction

```
ID   P32241|VIPR1_HUMAN
FT   SIGNAL        1     30
FT   REGION        1     11       N-REGION.
FT   REGION       12     23       H-REGION.
FT   REGION       24     30       C-REGION.
FT   TOPO_DOM     31    147       NON CYTOPLASMIC.
FT   TRANSMEM    148    167
FT   TOPO_DOM    168    178       CYTOPLASMIC.
FT   TRANSMEM    179    200
FT   TOPO_DOM    201    219       NON CYTOPLASMIC.
FT   TRANSMEM    220    243
FT   TOPO_DOM    244    254       CYTOPLASMIC.
FT   TRANSMEM    255    275
FT   TOPO_DOM    276    294       NON CYTOPLASMIC.
FT   TRANSMEM    295    319
FT   TOPO_DOM    320    338       CYTOPLASMIC.
FT   TRANSMEM    339    359
FT   TOPO_DOM    360    370       NON CYTOPLASMIC.
FT   TRANSMEM    371    391
FT   TOPO_DOM    392    457       CYTOPLASMIC.
//
```

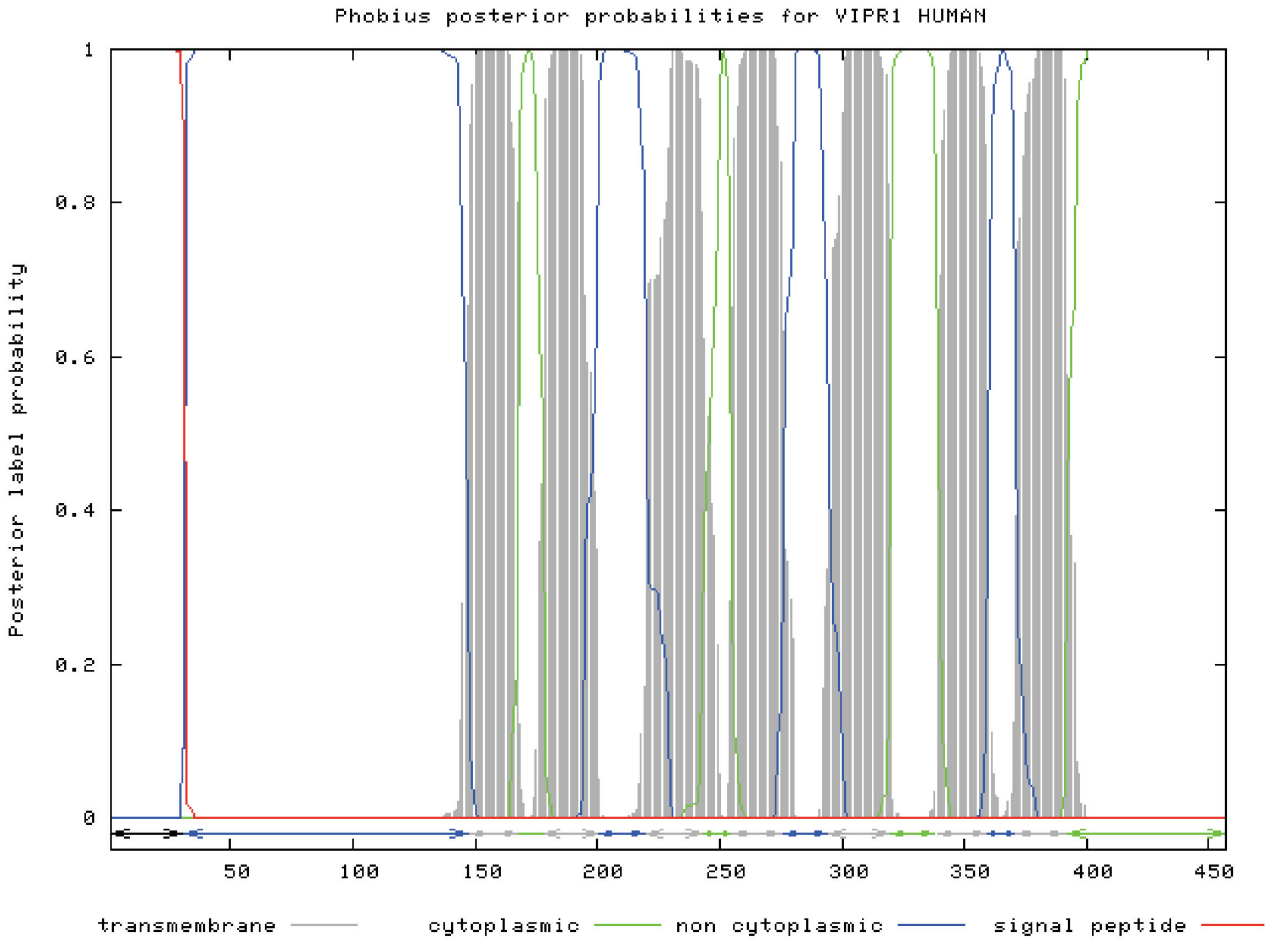

Figure 2. Output from the Phobius web server. An optional posterior probability plot is included in the prediction result.

specify such constraints for a prediction. The user may specify that a residue resides in a cytoplasmic loop, non-cytoplasmic loop or a transmembrane segment. One can also specify that the N-terminal part of the sequence is a signal peptide.

Here we maximize $P(Labels, Sequence|Model)$ $P(Labels|Constraints)$. This is implemented by a modification in the forward–backward (11) calculations; we multiply the forward probability for a state with the $P(Label|Constraint)$ in the constrained sequence positions.

As the membrane, signal peptide or a cytoplasmic loop states are uniquely identified by one single label in the Phobius model (Figure 1), we set $P(Label|Constraint)$ to 1 for the label corresponding to the constraint and 0 for all other labels in the constrained position. Non-cytoplasmic loops, on the other hand, can have two different labels. Here we assign 0.5 probability to each of the two constrained labels, and 0 to all other labels.

PREDICTION WITH HOMOLOGS

Since homologous sequences are likely to share both transmembrane topology and absence or presence of signal peptides, we can gain extra support for a prediction by examining the query sequence's homologs. This is the supporting idea for PolyPhobius, whose algorithm is described in a separate paper (12).

Here the server BLASTs the query sequence against UniProt. Hits with an E-value lower than 1E–5 covering more than 75% of the sequence length are used as support for the prediction. The full-length sequences are then realigned using a multiple sequence alignment program, and weighted with the Henikoff and Henikoff weighting scheme (13).

When we measured the performance of the approach, we found a significant increase in accuracy for transmembrane topology prediction accuracy (from 67.8 to 74.7% correct topologies) and as well as improvement in signal peptide prediction accuracy (increase in Matthews correlation from 0.901 to 0.921) as compared to Phobius without homology–enrichment (12).

The user can also submit his own alignment in Fasta format. In this case, the transmembrane topology and presence of signal peptide of the first sequence will be predicted taking the other sequences in the alignment into account.

IMPLEMENTATION

The Phobius web server is implemented as a Perl CGI-script. Plots are produced by gnuplot. Normal predictions are made with the ANHMM package (our unpublished data), while constrained predictions and predictions with homologs are done by HomologHMM package (12). Multiple sequence alignments are produced with Kalign (14).

AVAILABILITY

The Phobius web server is available at http://phobius.cgb.ki.se/ and http://phobius.binf.ku.dk/. Stand-alone versions of the software for academic users for Linux and SunOS are available on request.

ACKNOWLEDGEMENTS

Funding to pay the Open Access publication charges for this article was provided by Pharmacia corp.

Conflict of interest statement. None declared.

REFERENCES

1. Lao,D.M., Arai,M., Ikeda,M. and Shimizu,T. (2002) The presence of signal peptide significantly affects transmembrane topology prediction. *Bioinformatics*, **18**, 1562–1566.
2. Klee,E.W. and Ellis,L.B.M. (2005) Evaluating eukaryotic secreted protein prediction. *BMC Bioinformatics*, **6**, 256.
3. Käll,L., Krogh,A. and Sonnhammer,E.L.L. (2004) A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.*, **338**, 1027–1036.
4. Krogh,A., Larsson,B., vonHeijne,G. and Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
5. Nielsen,H., Engelbrecht,J., Brunak,S. and vonHeijne,G. (1997) A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Int. J. Neural Syst.*, **8**, 581–599.
6. Bendtsen,J.D., Nielsen,H., vonHeijne,G. and Brunak,S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.*, **340**, 783–795.
7. Schwartz,R. and Chow,Y. (1990) The N-best algorithm: an efficient and exact procedure for finding the N most likely sentence hypotheses. In *Proceedings of ICASSP 1990*, pp. 81–84.
8. Melen,K., Krogh,A. and vonHeijne,G. (2003) Reliability measures for membrane protein topology prediction algorithms. *J. Mol. Biol.*, **327**, 735–744.
9. Daley,D.O., Rapp,M., Granseth,E., Melen,K., Drew,D. and vonHeijne,G. (2005) Global topology analysis of the Escherichia coli inner membrane proteome. *Science*, **308**, 1321–1323.
10. Henricson,A., Käll,L. and Sonnhammer,E.L.L. (2005) A novel transmembrane topology of presenilin based on reconciling experimental and computational evidence. *FEBS J.*, **272**, 2727–2733.
11. Rabiner,L. (1989) A tutorial on hidden Markov models and selected aplications in speech recognition. *Proceedings of the IEEE*, **77**, 257–286.
12. Käll,L., Krogh,A. and Sonnhammer,E.L.L. (2005) An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics*, **21**(Suppl. 1), 251–257.
13. Henikoff,S. and Henikoff,J.G. (1994) Position-based sequence weights. *J. Mol. Biol.*, **243**, 574–578.
14. Lassmann,T. and Sonnhammer,E.L.L. (2005) Kalign–an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*, **6**, 298.